Marja Vierros

Linguistic Annotation of the Digital Papyrological Corpus: *Sematia*

1 Introduction: Why to annotate papyri linguistically?

Linguists who study historical languages usually find the methods of corpus linguistics exceptionally helpful. When the intuitions of native speakers are lacking, as is the case for historical languages, the corpora provide researchers with materials that replaces the intuitions on which the researchers of modern languages can rely. Using large corpora and computers to count and retrieve information also provides empirical back-up from actual language usage. In the case of ancient Greek, the corpus of literary texts (e.g. *Thesaurus Linguae Graecae* or the Greek and Roman Collection in the *Perseus Digital Library*) gives information on the Greek language as it was used in lyric poetry, epic, drama, and prose writing; all these literary genres had some artistic aims and therefore do not always describe language as it was used in normal communication. Ancient written texts rarely reflect the everyday language use, let alone speech. However, the corpus of documentary papyri gets close. The writers of the papyri vary between professionally trained scribes and some individuals who had only rudimentary writing skills. The text types also vary from official decrees and orders to small notes and receipts. What they have in common, though, is that they have been written for a specific, current need instead of trying to impress a specific audience. Documentary papyri represent everyday texts, utilitarian prose,[1] and in that respect, they provide us a very valuable source of language actually used by common people in everyday circumstances.

This significant text corpus is openly available to us in digital form. The *Papyrological Navigator* (PN)[2] hosts the *Duke Databank of Documentary Papyri* and provides a search engine as well. However, any deeper linguistic research cannot be performed. The search engine at PN is mainly designed for the needs of historians and editors of papyrus texts in locating parallels and sources using word-string searches. In order to utilize the text corpus linguistically, it needs to be enriched with linguistic information, i.e. it needs to be linguistically annotated.[3] Linguistic annotation can concern many different levels of language, usually morphology, syntax, semantics,

1 Cf. Wagner – Outhwaite – Beinhoff 2013, 4.
2 http://papyri.info/.
3 A very clear textbook on linguistic annotation and corpus linguistics in general is Kübler – Zinmeister 2015. See also e.g. Wynne 2005 on developing linguistic corpora.

https://doi.org/10.1515/9783110547450-006

or pragmatics. Even the basic morphological annotation alone can provide for complex linguistic queries. The literary Greek corpus has recently been automatically lemmatized and morphologically parsed.[4] Greek that is found in papyri deserves to be similarly treated so that the literary language can be compared with the utilitarian prose found in papyri, enabling our views on historical developments and variation of Greek language to be as full as they can be.

In this paper, I will discuss the criteria and approach which I have chosen while planning the *Sematia* corpus and platform.[5] While this is an ongoing process and plans often are subject to change, it is still worthwhile to explain what lies behind the selected approach, what the future plans are and possible new directions and, finally, what can be achieved with all this work.

2 Corpus design

One key factor in corpus design generally is that the corpus is representative. Whether we want a holistic or strictly selected corpus, depends on the research questions for which the corpus is meant to provide answers. If we want answers from a certain domain of texts (e.g. private letters), we select only those texts into the corpus. Similarly, whether we want a synchronic or diachronic corpus depends on whether we want to examine changes in language used within a certain time span or not. In historical linguistics, corpora are generally diachronic.

The papyrological corpus in PN is a growing and a changing one. It includes all published documentary papyri, and the Greek material ranges approximately from the IV century BC to the IX century AD. Newly published texts are added into the database by the academic community of papyrologists via the online *Papyrological Editor* (PE), where a board checks and votes on the submissions.[6] Also, mistakes (typos or wrong readings etc.) in the texts that already exist in the corpus, can be corrected via the same *Editor*. This is one reason for the idea that *Sematia* should also be kept open-ended, so that ideally it could include the whole corpus, which represents the Greek used in documentary papyri for a period of about a thousand years. Thus, at the moment, the corpus design is a loose one, but users (both the annotators and the researchers who only wish to perform queries) can decide on a case-by-case basis what they want to annotate or include in their searches. Once a version of a text has been annotated, that annotation is stable, but if the system alarms us that there has been a change introduced into the base text in the PN, the annotator (or someone else,

4 Celano 2017.

5 https://sematia.hum.helsinki.fi. I warmly thank the developer, Erik Henriksson, for all his ideas and efforts.

6 Sosin 2010, cf. also Reggiani 2017, 232–40.

for that matter) can renew the annotation on that text, if it seems warranted.[7] The process of getting texts annotated is slow at the moment, since it is performed semi-automatically (more on this aspect below). The choice of texts to be annotated is not authoritatively dictated by us; the choice is made by the users, so anyone wanting to have a specifically chosen set of material, can proceed in annotating the papyri. This way s/he also makes a contribution towards the annotation of the whole corpus. And when there are more texts already annotated, each researcher may select his/her own subcorpus and perform queries only on them (either in the *Sematia* platform or after downloading all the selected annotations for external use). The latter option makes the research more easily replicable (a basic requirement in corpus linguistic research).

Corpus design also includes deciding over the level of annotation and what features are annotated and how. At the moment, our basic approach is to include the morphological and syntactic annotation in the form of dependency treebanks. We follow the *Ancient Greek and Latin Dependency Treebank* system.[8] *Sematia* is designed to provide a 'basic' level of annotation, because we have this holistic idea of the whole corpus eventually being annotated; the research questions must not in this case be strictly decided beforehand. However, since the automatic morphological parsing has been performed on literary texts as mentioned above, this is a logical next step for the whole papyrological text corpus as well. This, in turn, would make the manual syntactic treebanking somewhat quicker, as the morphological forms would be more accurate than they are now to begin with (on the process of annotation, more detailed description below).

3 How to annotate papyri?

Why should we devote a section on how to linguistically annotate papyrus texts? Because the papyri represent ancient textual material often preserved in a fragmentary condition. The organic writing material has suffered damage of many kinds. But, due to the importance of papyri as a source, papyrologists work very hard on reading, transcribing, and reconstructing them, i.e. editing the text, so that other researchers can also use that source. Still, many gaps and question marks can remain in the editions. All this is encoded within the text in the digital edition, in TEI EpiDoc XML,[9] and for this reason we do not have simple access to the raw text that could simply be uploaded for some linguistic annotation tool. In fact, the editorial work gives us

7 This type of alarm system has not yet been established, but it is on our agenda.

8 https://perseusdl.github.io/treebank_data; Bamman – Crane 2011.

9 https://sourceforge.net/p/epidoc/wiki/Home.

plenty of material that we can and should also use in the linguistically annotated corpus. Therefore, we need to preprocess the texts available in the PN in a certain way.

3.1 Preprocessing

The *Sematia* tool was first developed mainly for the above-mentioned preprocessing need. It creates two parallel layers of the same text; one being a sort of diplomatic edition (called "original"), and the other including the editorial suggestions (called "standard"). The tool has already been described in another article,[10] thus I will not present the details here. What makes the *Sematia* corpus special, is that *both* of these parallel layers are linguistically annotated. This way it is possible to study only the version that has truly been preserved for us (the original layer), or to compare the actual preserved text with its standardized version. The differences in this comparison can be turned into a third layer (called "variation"), which I will briefly discuss later.

3.2 Annotation

In order for this corpus to be beneficial for all Greek linguists, I decided that we should follow the same scheme and standard used in the corpora of ancient Greek. This means the *Ancient Greek and Latin Dependency Treebank* that includes Greek literature. In addition, the PROIEL treebank (New Testament and some Greek prose) follows the Dependency Grammar.[11] In the annotation of papyri, we follow the Guidelines of AGDT.[12] At the moment, we use the external annotation environment, *Arethusa*, provided by the *Perseids* Platform,[13] with which we have an API integration in *Sematia*. This means that a text can be exported directly from *Sematia* to *Perseids* and *Arethusa*, and after it has been annotated, a member of the *Sematia* board (at the moment the project director) goes through the annotations in *Perseids* and either accepts or returns them to the annotator to be corrected. After the approval, the treebanks are committed back to *Sematia* (both the *GitHub* repository and the online site).

The process of annotation in *Arethusa* includes the tokenization; i.e. tokenization is done when the plain text is imported into *Arethusa*, not into *Sematia*. The text receives an automatic lemmatization and morphological tagging (by *Morpheus*). But all

10 Vierros – Henriksson 2017.

11 Both treebanks have also been modified for the Universal Dependencies site (http://universaldependencies.org), where they can be accessed together with many other languages.

12 Version 1.1: Bamman – Crane 2008, version 2.0: Celano 2014. Version 2.0 is to be followed, but version 1.1 has sometimes more useful examples and more detailed explanations.

13 http://sites.tufts.edu/perseids.

the lemmas and morphological tags need to be checked and corrected by the human annotator; there are several forms and lemmas in the papyri, which *Morpheus*, being designed for classical Greek, does not recognize, for example the Egyptian names. Moreover, *Morpheus* does not do well in selecting the correct form from several homonyms. The syntactic annotation and dependencies have to be performed manually by the annotator. In other words, using *Arethusa* is convenient up to a point; but it is also quite laborious and thus expensive as it needs human resources: skilled annotators and their time. Nevertheless, in the end, we do get accurate annotations that can most likely be used in training automatic syntactic and morphological parsers in the future.

The process can be presented by an example with images. Our sample sentence is the second sentence of a letter from Petenephotes to Valerius, written on a potsherd in the garrison of Mons Claudianus in the Eastern Desert (O.Claud. II 245,2–7; mid II century AD):

> [1][καλῶς] |[3] πυήσις, ἄδελφε, ἐὰ[ν ἔλθῃ] |[4] ἡ πορήε τῇ νυκτὶ ταύτῃ \ πέμψας μοι / |[5] τρία ζεύγη ἄρτων ἐπὶ οὐκ ἐ|[6]χο ἄρτους καὶ ὅταν ἔλθῃ ἡ πο|[7]ρήα πέμψω συ αὐτά.
>
> 3. *l.* ποιήσεις 4. *l.* πορεία 5–6. *l.* ἔ|χω 6–7. *l.* πο|ρεία 7. *l.* σοι
>
> *Please, brother, if the caravan arrives tonight, send me three pairs of bread as I do not have any bread and when the caravan arrives I shall send them to you.*

Note that the apparatus has several corrections (standardizations), but not for the ι/ει confusion in the conjunction ἐπὶ (l. ἐπεί), l. 5. This is the standard practice in this edition. Other so-called orthographic mistakes are usually standardized in the apparatus, but not the most common one between ι / ει, because the editors apparently consider this such a common, parallel variant that it can no longer be considered as a 'mistake' (see also the chapter by J. Stolk in this volume for problems that this type of fluidity between editorial corrections can cause).

The standard and the original layers of this sentence in the *Arethusa* treebank tool are presented in Figg. 1 and 2. Only the syntactic trees can be seen in the screenshots, and only one lemma/morphological analysis (that of the highlighted word), in this case the conjunction mentioned above. This is emphasized here, because an automatic parser would automatically take this word as the preposition ἐπὶ, but when the human annotator checks the sentence, s/he notices that the preposition is not the correct interpretation, and can make the necessary correction, even though the word is not editorially corrected in the original electronic source of ours, in the PN.

The differences between the layers are apparent in the images; the supplied text, for example, is not annotated in the original layer, it is represented with a dummy marker SU so that the annotator notices that something is missing there and the supplemented word does not end up in the corpus of original layers. This also leaves some of the branches of the sentence tree hanging in the air, as some words that

would be the heads on which other words depend on, are not preserved in the papyrus. The non-standard orthography in the original will not prevent the annotator from recognizing and marking correct lemmata for the forms, thus lemma searches will find all variant spellings of the words from the original layers.

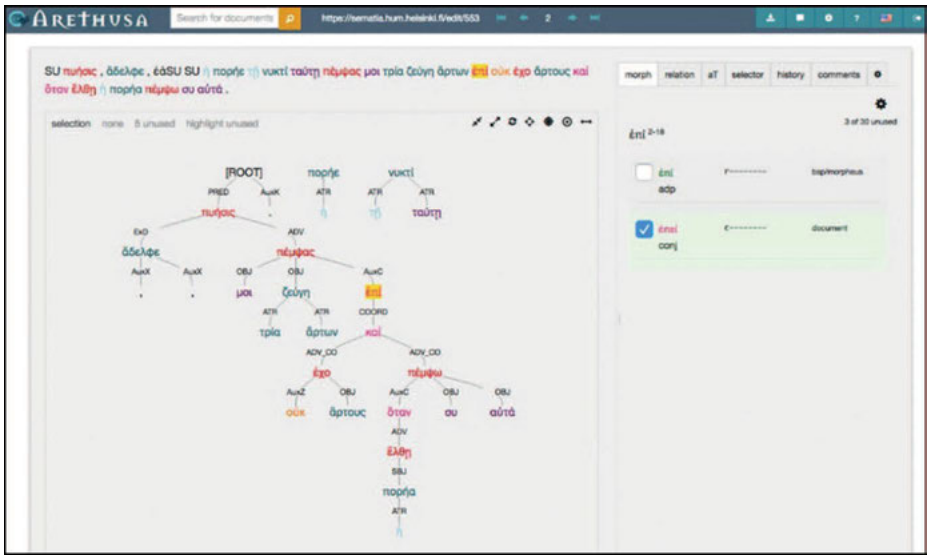


Fig. 1: Original layer of the sentence [1] in *Arethusa*.

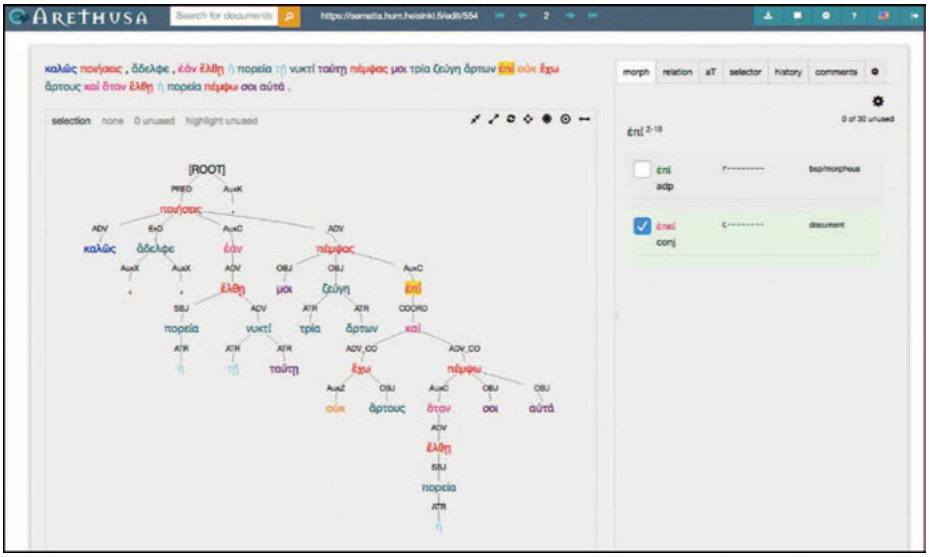


Fig. 2: Standard layer of the sentence [1] in *Arethusa*.

The underlying XML forms show us what the whole annotation entails. Fig. 3 presents the XML of the original layer's annotation of the same sentence [1].

```
<sentence document_id="https://sematia.hum.helsinki.fi/edit/553" id="2" span="" subdoc="">
   <word form="SU" head="" id="1" lemma="" postag="" relation="" />
   <word form="πυήσις" head="0" id="2" lemma="ποιέω" postag="v2sfia---" relation="PRED" />
   <word form="," head="4" id="3" lemma="punc1" postag="u--------" relation="AuxX" />
   <word form="ἄδελφε" head="2" id="4" lemma="ἀδελφός" postag="n-s---mv-" relation="ExD" />
   <word form="," head="4" id="5" lemma="punc1" postag="u--------" relation="AuxX" />
   <word form="ἐὰSU" head="" id="6" lemma="" postag="" relation="" />
   <word form="SU" head="" id="7" lemma="" postag="" relation="" />
   <word form="ἡ" head="9" id="8" lemma="ὁ" postag="l-s---fn-" relation="ATR" />
   <word form="πορήε" head="" id="9" lemma="πορεία" postag="n-s---fn-" relation="" />
   <word form="τῇ" head="11" id="10" lemma="ὁ" postag="l-s---fd-" relation="ATR" />
   <word form="νυκτὶ" head="" id="11" lemma="νύξ" postag="n-s---fd-" relation="" />
   <word form="ταύτῃ" head="11" id="12" lemma="οὗτος" postag="p-s---fd-" relation="ATR" />
   <word form="πέμψας" head="2" id="13" lemma="πέμπω" postag="v-sapamn-" relation="ADV" />
   <word form="μοι" head="13" id="14" lemma="ἐγώ" postag="p1s---md-" relation="OBJ" />
   <word form="τρία" head="16" id="15" lemma="τρία" postag="n-p---na-" relation="ATR" />
   <word form="ζεύγη" head="13" id="16" lemma="ζεῦγος" postag="n-p---na-" relation="OBJ" />
   <word form="ἄρτων" head="16" id="17" lemma="ἄρτος" postag="n-p---mg-" relation="ATR" />
   <word form="ἐπὶ" head="13" id="18" lemma="ἐπεί" postag="c--------" relation="AuxC" />
   <word form="οὐκ" head="20" id="19" lemma="οὐ" postag="d--------" relation="AuxZ" />
   <word form="ἐχο" head="22" id="20" lemma="ἔχω" postag="v1spia---" relation="ADV_CO" />
   <word form="ἄρτους" head="20" id="21" lemma="ἄρτος" postag="n-p---ma-" relation="OBJ" />
   <word form="καὶ" head="18" id="22" lemma="καί" postag="c--------" relation="COORD" />
   <word form="ὅταν" head="27" id="23" lemma="ὅταν" postag="c--------" relation="AuxC" />
   <word form="ἔλθῃ" head="23" id="24" lemma="ἔρχομαι" postag="v3sasa---" relation="ADV" />
   <word form="ἡ" head="26" id="25" lemma="ὁ" postag="l-s---fn-" relation="ATR" />
   <word form="πορήα" head="24" id="26" lemma="πορεία" postag="n-s---fn-" relation="SBJ" />
   <word form="πέμψω" head="22" id="27" lemma="πέμπω" postag="v1sfia---" relation="ADV_CO" />
   <word form="συ" head="27" id="28" lemma="σύ" postag="p-s---mn-" relation="OBJ" />
   <word form="αὐτά" head="27" id="29" lemma="αὐτός" postag="p-p---na-" relation="OBJ" />
   <word form="." head="0" id="30" lemma="punc1" postag="u--------" relation="AuxK" />
</sentence>
```

Fig. 3: XML of the treebanked original layer of the sentence [1].

We can see that the annotation includes the existing form of the word in the sentence, its head, its lemma, the postag and the syntactic relation of the word in the sentence. The postag includes the whole morphological analysis: part-of-speech, person, number, tense, mood, voice, gender, case and degree. For example, the form πεμψας (word id 13) is a verb, singular, aorist participle active in the masculine nominative. The postag gives the very basic morphological analysis, and we could occasionally hope for something more specific, such as distinguishing proper nouns from common nouns or possessive pronouns from other pronouns, but as of this moment, *Morpheus* gives us these. In the future, other automatic parsers might take these distinctions into account more easily. However, even this morphological analysis enables us to search complex linguistic structures, especially when combined with the lemma and syntactic annotations. This, I think, is sufficient to fulfill the need of basic linguistic annotation for the *Sematia* corpus. Other levels of annotation, e.g. semantic or information structure annotations, would take considerably more time and effort.

4 Metadata and its purposes

The date and place of origin of each text are vital when we wish to see in which time periods and in which areas certain linguistic features appear. They are generally provided in the papyrus editions and presented also in the PN metadata field, from whence they are automatically drawn into *Sematia*.

As mentioned already in Vierros – Henriksson 2017, we add some metadata, which is not available in PN, namely aspects relating to the handwriting and the writers vs. authors. Some changes have been planned for these metadata fields and they will be implemented in the near future. The purpose is to identify text parts written in the same hand. When imported to *Sematia*, each document is divided into 'acts of writing' by the element `<handShift>`, i.e. each section written by a different writer receives its own layers and treebanks. Since there are often papyrus archives in which the same hand can have written several documents, it is important to link these acts of writing together, so that we can also try to study idiolects and compare certain writers to others. At the moment of writing, we can add metadata concerning the handwriting[14] and concerning the writer, the author and the addressee.[15] See Fig. 4 for an example on the metadata in O.Claud. II 245 (which only has one hand). In many cases, however, the name of the actual writer is not known, e.g. in private letters the sender of the letter is taken as the author, but the actual writer is not necessarily the same person as the author, nor is he named. In contracts, the names of the contracting parties are mentioned, but the scribe who draws up the text or who pens down the letters onto the papyrus often remains unnamed. Therefore, the *Trismegistos People* ID cannot be used in identifying the hands, since we have so many hands without names to connect them with. Our intention is to give each hand an ID of its own. The hands that have been identified to come from one writer (sometimes a very difficult task), can be connected to the same ID. The hand-ID will make the current metadata field "Same hand" obsolete.[16]

For the purposes of studying linguistic register and features typical of certain text types, we have also included the fields in which we can insert metadata on the text type and the addressee.

14 There are fields for the description of the handwriting in the edition or some other scholarly source, the description of the handwriting by the annotator, and the "Same hand" field, i.e. list of other documents, where the same hand is said to appear. These fields are text-based, and thus they do not provide good searchable data. Every papyrologist is also well aware of the lack of precision of these descriptions in different editions.

15 For each person the annotator can add the name, title and the *Trismegistos People* ID (http://www.trismegistos.org/ref/index.php).

16 In its current state, the field is not very usable, user-friendly or accurate; the list of other documents where the same hand appears is done in stable URLs of the documents in PN, but one document can contain several hands.

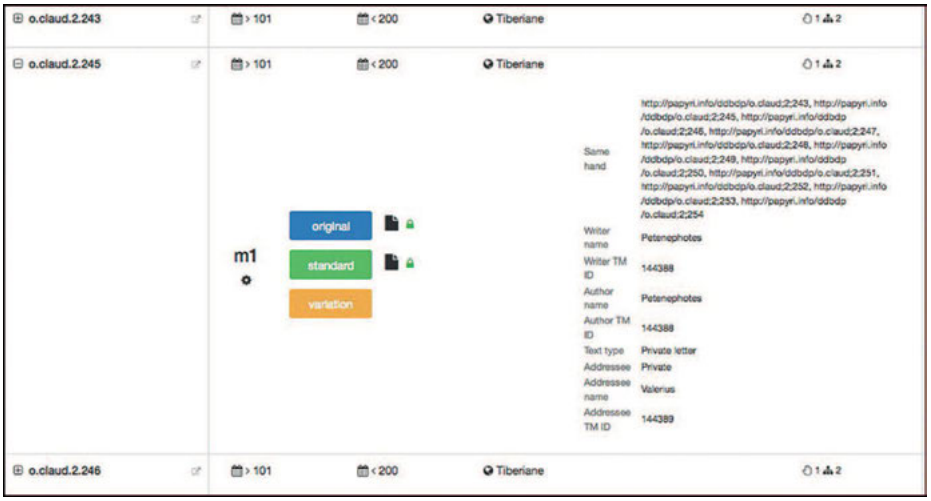


Fig. 4: Screenshot of the main view from *Sematia*, when the document O.Claud. II 245 is expanded (but O.Claud. II 243 and 246 are not). On the right, the metadata inserted in *Sematia* by the annotator is visible. The field "Same hand" is extensive with many documents also written in Petenephotes's hand. The editor mentions that the writer is Petenephotes himself,[17] thus he is both the author and the writer. Clicking from the blue "original" or green "standard" buttons would take you to the text, and clicking the paper icon next to those buttons, you could view the treebank XML.

5 Sample results, i.e. what queries can find

The treebank XML files (including the metadata) in *Sematia* can be exported for querying in external treebank query tools.[18] It is possible to export the treebanks of all layers together, or choose the original or standard layers separately. I will not go through all the possibilities the external search engines can give for linguists;[19] I will describe some sample searches that can be performed on the *Sematia* site itself.[20] There, too, it is possible to search only from the treebanks of the original layers or only from the standard layers, but one of the essential features is the possibility to find instances where the original and standard layers differ. This is where we can get

17 Bülow-Jacobsen 1997, 69.

18 E.g. *SETS Treebank Search*, *PML Tree Query Engine* or *XQuery/BaseX*, cf. Vierros – Henriksson 2017, 13.

19 One thorough treebank-based study on ancient languages is Korkiakangas 2016, in which the author has been able to study under which conditions the Latin accusative began to be used as the subject case in VIII and IX centuries.

20 https://sematia.hum.helsinki.fi/tools.

more deeply into linguistic variation. For example, it is very simple to search for instances where one grammatical case is used when editors have thought that a different case would have been more understandable, or more standard (what the editorial standardizations might have meant in different times when papyri have been edited, see the chapter by J. Stolk in this volume). The search fields in *Sematia* employ Regular Expressions (regex). The searches can naturally be limited in multiple ways, either by metadata fields or by the other field related to linguistic annotation, e.g. searching only objects or subjects, or only verbs or pronouns. More complex searches combining several words or forms would need to be made externally.

An example search concerning the grammatical case, the dative instead of the genitive, is presented in Fig. 5. Since the postag holds the case in the 8th place of the string, we can use the values for dative (d) in the original layer's postag field and genitive (g) in the 8th place in the standard layer's field, and let other places of the string be whatever else by using the wildcard (.); the beginning of the string is marked by (^). The values (d) and (g) can have different meanings in other positions in the postag, thus it is good to define the exact location. In other words, when using the search, it is vital to know how the annotations have been made, i.e. what each symbol means e.g. in the postag field. The guidelines of annotation need to be known and understood.

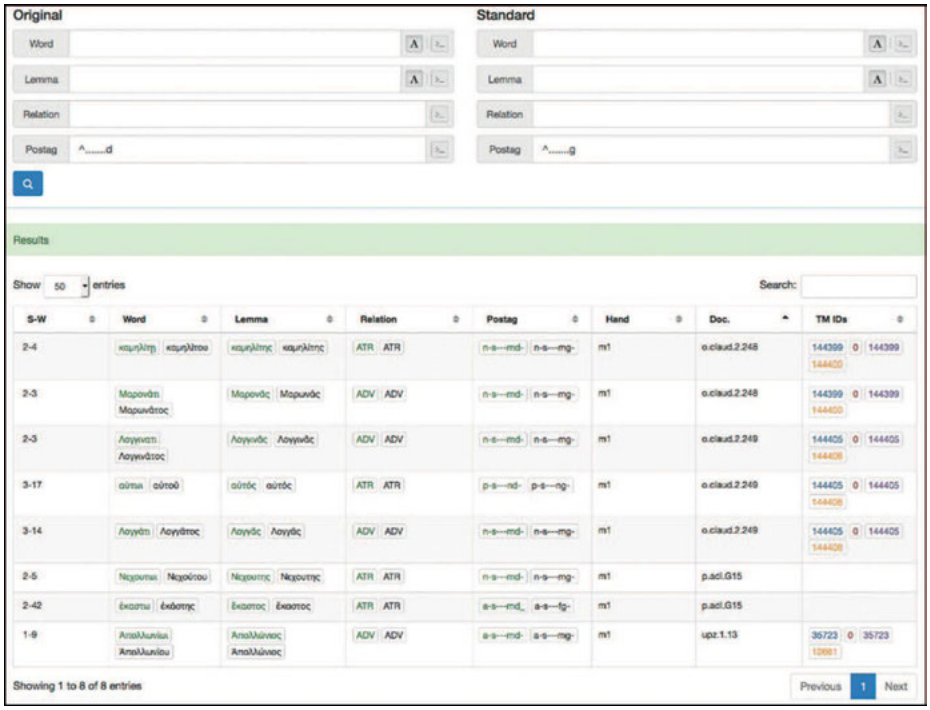


Fig. 5: A screenshot of the search and results in *Sematia* for the dative in the original layer vs. the genitive in the standard layer.

The search gives eight results with the limited data we have in *Sematia* at the moment (2017, ca. 100 papyri). The result list can be ordered according to different fields, in Fig. 5 it is ordered by the document name. We can see that some of the instances may, in fact, signal orthographic confusion based on phonological variation rather than case confusion (e.g. Νεχουτωι / Νεχουτου),[21] but some of the instances more clearly tell that the writer has, for some reason, really chosen the dative rather than the expected genitive (e.g. Μαρονατι / Μαρονατος). Similarly, we could bring up e.g. all prepositions in the texts by simple postag query (^r), or see where singular verb forms appear instead of plural verb form (^v.s vs. ^v.p). In the latter search, the results again point to the interplay of phonological factors confusing the morphological interpretations. See Fig. 6, where two out of three of the singular vs. plural verb form are forms consisting of graphemes αι / ε, both marking the phoneme /e/ at this time, and the third one has α / ε confusion, which was also perhaps due to weak pronunciation of the unstressed vowel. These results give us material for further research on phonology playing a part in the morphological mergers in Greek, and the impact of education in writers' ability or inability to use standard orthography in such occasions, but they also provide us with material for enhancing our tools in the future.

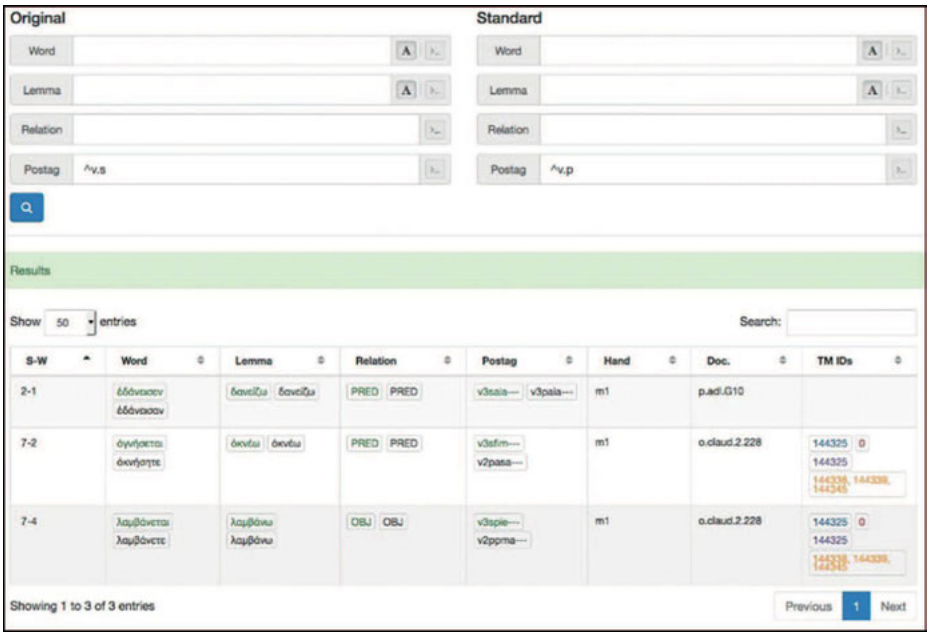


Fig. 6: A screenshot of the search in *Sematia* for a singular verb in the original layer vs. a plural verb in the standard layer.

21 See, however, DAHLGREN 2017, 90 ff. on phonological variation of /o, u/ possibly playing a role in case variation.

6 Future plans

A variation layer has been on our agenda since the beginning and it was discussed already in the previous article to some extent.[22] With the above described method of comparison between the original and standard layers, we can only find variation (instances where there really are differences between the layers), when there is a regularization in the PN, or when the annotator has marked these differences in the treebank XML after seeing a difference not available in the PN version. These comparisons and differences are planned to be automatically retrieved into *Sematia* to form the basis for the variation layer. In addition, we do need a way to manually encode other types of linguistic variation in this layer for several reasons. For example, there is a need to further specify certain differences as more phonological or more morphological in nature. Secondly, some variation is impossible to detect from the annotations when the postag does not really describe what we have in the text. I will give an example of this type of case with one sentence from a letter written by Ammonius to Apollonius (O.Claud. I 155,3–5; II century AD):

[2] Ἁρπαήσιος ὁ κιβαριάτης εἶ|[4]ρηκέ μοι ὅτι ἐπιστολὴν ἔλα|[5]βα ἀπὸ τῆς γυναικός μου.

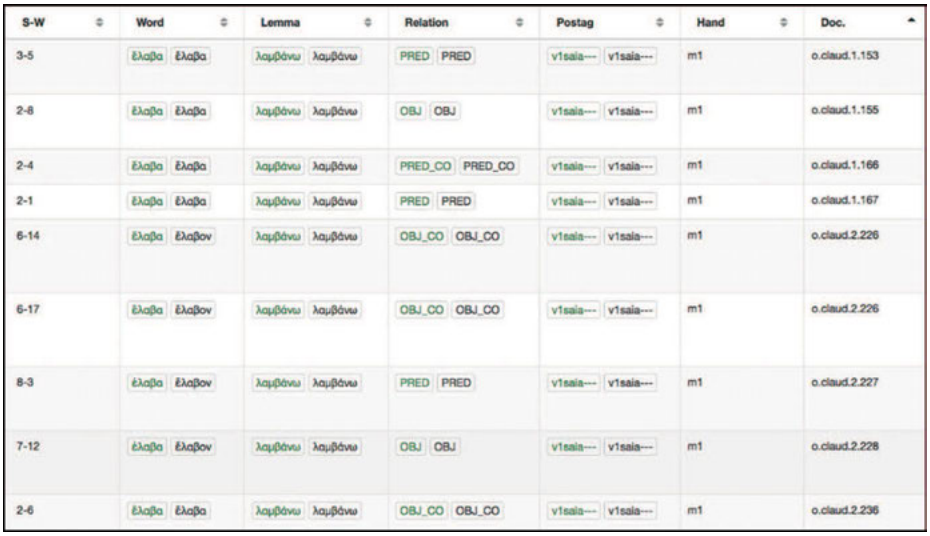
Harpaesius, the cibariator*, has told me that I have got a letter from my wife.*

The form ἔλαβα, "I got", has not been corrected in the apparatus, even though it represents mixed morphology; the aorist of the verb λαμβάνω would be ἔλαβον according to the classical standard (the second i.e. 'strong' aorist), but in the *Koine* the athematic endings of the first i.e. 'weak' aorist (-α for the first person) were occasionally used (and they are the ones used in modern Greek).[23] In the Mons Claudianus *ostraka* so far annotated in *Sematia*, there are nine attestations of the form ἔλαβα (plus three times written as αἴλαβα),[24] but the editor has fluctuated in correcting it in the apparatus (see Fig. 7). We can find this word by using the word search, but as can be seen from the postag, it is not possible to indicate this type of variation there; the postag is the same in both ἔλαβα and ἔλαβον: first person singular aorist form. It would be very convenient to mark this up in the separate variation layer as mixed morphological endings in the aorist.

22 Vierros – Henriksson 2017, 13.

23 Cf. Horrocks 2010, 109–10 and 143–4 on the developments of past-tense morphology.

24 All three in O.Claud. II 236.



S-W	Word	Lemma	Relation	Postag	Hand	Doc.
3-5	ἔλαβα ἔλαβα	λαμβάνω λαμβάνω	PRED PRED	v1saia--- v1saia---	m1	o.claud.1.153
2-8	ἔλαβα ἔλαβα	λαμβάνω λαμβάνω	OBJ OBJ	v1saia--- v1saia---	m1	o.claud.1.155
2-4	ἔλαβα ἔλαβα	λαμβάνω λαμβάνω	PRED_CO PRED_CO	v1saia--- v1saia---	m1	o.claud.1.166
2-1	ἔλαβα ἔλαβα	λαμβάνω λαμβάνω	PRED PRED	v1saia--- v1saia---	m1	o.claud.1.167
6-14	ἔλαβα ἔλαβον	λαμβάνω λαμβάνω	OBJ_CO OBJ_CO	v1saia--- v1saia---	m1	o.claud.2.226
6-17	ἔλαβα ἔλαβον	λαμβάνω λαμβάνω	OBJ_CO OBJ_CO	v1saia--- v1saia---	m1	o.claud.2.226
8-3	ἔλαβα ἔλαβον	λαμβάνω λαμβάνω	PRED PRED	v1saia--- v1saia---	m1	o.claud.2.227
7-12	ἔλαβα ἔλαβον	λαμβάνω λαμβάνω	OBJ OBJ	v1saia--- v1saia---	m1	o.claud.2.228
2-6	ἔλαβα ἔλαβα	λαμβάνω λαμβάνω	OBJ_CO OBJ_CO	v1saia--- v1saia---	m1	o.claud.2.236

Fig. 7: A screenshot of the search results in *Sematia* for the word form 'ελαβα'. In the "word" column, the words in green come from the "original" layers and the words in black come from the "standard" layers. In O.Claud. volume I, the form was not standardised according to the classical norm, whereas in volume II it was (with one exception).

We will be developing *Sematia* and similar tools further.[25] One idea is to have the whole papyrological corpus already present in *Sematia*, and updated in set intervals, i.e. there would no longer be the need to import texts individually. Phonological searches will be enabled on the whole corpus. We also aim at developing an automatic morphological parser for Greek found in papyri, with more accurate analysis than what *Morpheus* currently has.

25 The project "Digital Grammar of Greek Documentary Papyri" (ERC Starting Grant 2017 no. 758481) will use and develop these tools.

7 Bibliography

Bamman, D. – Crane, G. (2008), *Guidelines for the Syntactic Annotation of the Ancient Greek Dependency Treebank* 1.1, The Perseus Project, Tufts University, URL: http://nlp.perseus.tufts.edu/syntax/treebank/greekguidelines.pdf.

Bamman, D. – Crane, G. (2011), *The Ancient Greek and Latin Dependency Treebanks*, in *Language Technology for Cultural Heritage*, Berlin–Heidelberg, 79–98.

Bamman, D. – Mambrini, F. – Crane, G. (2009), *An Ownership Model of Annotation: The Ancient Greek Dependency Treebank,* in *Proceedings of the 8th Workshop on Treebanks and Linguistic Theories (TLT8),* at URL: http://www.perseus.tufts.edu/~ababeu/tlt8.pdf.

Bülow-Jacobsen, A. (1997), *The Correspondence of Petenephotes (243–254),* in *Mons Claudianus Ostraca Graeca et Latina II*, ed. by J. Bingen, A. Bülow-Jacobsen, W.E.H. Cockle, H. Cuvigny, F. Kayser, and W. Van Rengen, Le Caire.

Celano, G.G.A. (2014), *Guidelines for the Annotation of the Ancient Greek Dependency Treebank 2.0.*, URL: https://github.com/PerseusDL/treebank_data/edit/master/AGDT2/guidelines.

Celano, G.G.A. (2017), *Lemmatized Ancient Greek Texts*, URL: https://github.com/gcelano/LemmatizedAncientGreekXML.

Celano, G.G.A. – Crane, G. – Majidi, S. (2016), *Part of Speech Tagging for Ancient Greek*, "Open Linguistics" 2, 393–9.

Horrocks, G. (2010), *Greek. A History of the Language and Its Speakers. Second Edition*, Malden – Oxford – Chichester.

Korkiakangas, T. (2016), *Subject Case in the Latin of Tuscan Charters of the 8th and the 9th Centuries.* Helsinki.

Kübler, S. – Zinmeister, H. (2015), *Corpus Linguistics and Linguistically Annotated Corpora*, London – New York.

Reggiani, N. (2017), *Digital Papyrology I. Methods, Tools and Trends,* Berlin – Boston.

Sosin, J. (2010), *Digital Papyrology*, URL: http://www.stoa.org/archives/1263.

Vierros, M. – Henriksson, E. (2017), *Preprocessing Greek Papyri for Linguistic Annotation*, in *Journal of Data Mining and Digital Humanities. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*, ed. by M. Büchler and L. Mellerin, URL: http://jdmdh.episciences.org/paper/view/id/1385. (Version 1 was published in 2016)

Wagner, E.-M. – Outhwaite, B. – Beinhoff, B. (2013), eds., *Scribes as Agents of Language Change,* Boston.

Wynne, M. (2005), ed., *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford, URL: http://ota.ox.ac.uk/documents/creating/dlc.